Estimation Without Representation: Early Severe Acute Respiratory Syndrome Coronavirus 2 Seroprevalence Studies and the Path Forward

Bonnie E. Shook-Sa,[1] Ross M. Boyce,[2] and Allison E. Aiello[3,4]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [2]Division of Infectious Diseases, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [3]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA, [4]Department of Epidemiology, Carolina Population Center, University of North Carolina, Chapel Hill, North Carolina, USA

The recent development and regulatory approval of a variety of serological assays indicating the presence of antibodies against severe acute respiratory syndrome coronavirus 2 has led to rapid and widespread implementation of seroprevalence studies. Accurate estimates of seroprevalence are needed to model transmission dynamics and estimate mortality rates. Furthermore, seroprevalence levels in a population help guide policy surrounding reopening efforts. The literature to date has focused heavily on issues surrounding the quality of seroprevalence tests and less on the sampling methods that ultimately drive the representativeness of resulting estimates. Seroprevalence studies based on convenience samples are being reported widely and extrapolated to larger populations for the estimation of total coronavirus disease 2019 (COVID-19) infections, comparisons of prevalence across geographic regions, and estimation of mortality rates. In this viewpoint, we discuss the pitfalls that can arise with the use of convenience samples and offer guidance for moving towards more representative and timely population estimates of COVID-19 seroprevalence.

Keywords. address-based sampling; COVID-19; convenience sampling; seroprevalence; transmission.

LIMITATIONS TO THE GENERALIZABILITY OF EARLY SEVERE ACUTE RESPIRATORY SYNDROME CORONAVIRUS 2 SEROPREVALENCE STUDIES

In addition to direct health impacts, coronavirus disease 2019 (COVID-19) has caused an unprecedented level of disruption to social networks and economic systems. Phased "re-opening" policies are being guided by surrogate measures of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission, including symptom-based (ie, syndromic surveillance), test-based (ie, positivity rates), and facility-based (ie, hospitalizations) measures of disease activity [1]. The United States has experienced shortages of critical testing supplies, which has contributed to an underreporting of cases and struggling mitigation efforts [2]. Given the persistent issues, many individuals face accessing testing and the high proportion of subclinical infections that do not prompt care-seeking [3, 4], these metrics are suboptimal measures of disease activity in the community.

The recent development and rapid regulatory approval of serological assays indicating the presence of antibodies against SARS-CoV-2 has led to widespread implementation of seroprevalence studies [5]. These studies have used a wide range of assays and recruitment methods. Although issues surrounding test performance have been the focus of much debate [6], there has been much less discussion around the rigor and appropriateness of sampling frames. In this study, we illustrate some of the pitfalls associated with the use of convenience samples and provide guidance on best practices for quickly generating population-based estimates of seroprevalence.

To date, several large seroprevalence studies have been published in both the preprint and peer-reviewed literature. The majority of these studies have used convenience sampling, with participants recruited from online platforms (eg, Facebook), healthcare facilities, market research databases, or shopping centers [7–10]. The benefit of convenience sampling is that recruitment can occur relatively quickly and it is generally less expensive than probability-based sampling approaches. The major disadvantage is that resulting estimates will often not reflect the true seroprevalence in the underlying population due to selection bias. That is, recruitment methods and inherent factors that drive participation in convenience samples often lead to samples that do not reflect the underlying population in terms of demographic composition and risk factors for COVID-19 infection. Moreover, it is very difficult to estimate the level of bias introduced by convenience sampling, especially when there are few population-based studies available for comparison. This can make

Received 16 June 2020; editorial decision 10 July 2020; accepted 11 July 2020; published online July 16, 2020.

Correspondence: Bonnie E. Shook-Sa, DrPH, Department of Biostatistics, University of North Carolina at Chapel Hill, 135 Dauer Drive, 3101 McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420 (bshooksa@live.unc.edu).

DOI: 10.1093/infdis/jiaa429

extrapolation to underlying populations and estimation of mortality rates problematic.

Selection bias inherent in convenience samples is largely a result of competing factors that may influence an individual's participation. For example, an individual with a prior COVID-19 infection might be more likely to volunteer for a seroprevalence study due to recent symptoms compared with someone without a prior infection who has been asymptomatic, thus leading to an overestimate of seroprevalence in the underlying population. Likewise, persons in shopping centers and other public areas where recruitment occurs may be at higher risk for COVID-19 than the general population. In contrast, persons who are avoiding shopping centers, perhaps due to underlying illness or the presence of high-risk individuals in the household, are essentially excluded from participating in seroprevalence studies with this recruitment method. Other design features of convenience samples can result in underestimating seroprevalence. Studies based on social media recruitment have reported underrepresentation of older persons and overrepresentation of non-Hispanic whites [8]. However, African American and Hispanic populations have been disproportionally impacted by the COVID-19 pandemic [11]. Thus, underrepresentation in studies may not only contribute to underestimates of community seroprevalence but also to poorly targeted policy. Although weighting methods [8] or modeling approaches [12] can be used to improve representativeness of convenience samples, such methods rely on assumptions that cannot be validated. These methods typically assume that participants are like a random sample from the population stratified by a known set of characteristics. When this assumption holds, these methods provide unbiased estimates. Violations occur when participation is driven by the outcome itself or when participation and risk for the outcome are driven by factors not accounted for in the analysis. This leaves researchers to speculate about how these errors offset one another and if the results are truly representative [8, 10].

To illustrate some of these biases, we examined characteristics of the underlying population for a seroprevalence study conducted in Santa Clara County, California in early April [8]. Participants were recruited using targeted Facebook advertisements and community listserves. Although researchers tried to recruit such that the distribution of participants would accurately reflect the population geographically, persons in wealthier areas were overrepresented. The sample also underrepresented men, persons 65 and older, Hispanics, and Asians. Researchers weighted the sample such that the weighted distributions of participants by ZIP Code, sex, and race/ethnicity would reflect known county demographics, and weighted estimates of seroprevalence were produced. This approach assumes that participants in the study are like a random sample of Santa Clara residents stratified by ZIP Code, sex, and race/ethnicity. There is some evidence to question the validity of this assumption. Age, a factor correlated with COVID-19 risk [13], was not included as a weighting variable. Although 12.9% of Santa Clara residents are reported as being aged 65 or older, only 4.5% of the weighted sample represented this age group. Furthermore, characteristics such as occupation [14] and social distancing practices [15] are likely drivers of COVID-19 infection and were not accounted for in the weighting or analysis. The results of the Santa Clara study and other convenience samples have been publicized widely in the media [5, 7, 16], and thus the estimates from these studies have the potential to influence policymakers.

BEST PRACTICES FOR GENERATING A REPRESENTATIVE SAMPLING FRAME

Classic sample surveys achieve representation by ensuring that (1) all members of the population of interest have a chance of being included in the study, (2) members of the population are randomly selected for participation, and (3) researchers can quantify the chance that each sampled person was selected. These criteria are grounded in probability theory and have long been used to provide valid inference about target populations from concrete sampling frames, which are lists of population members from which samples are selected. Although sample surveys can suffer from generalizability concerns when there are problems with the sampling frame or participation, these errors have long been recognized, and methods have been developed to minimize errors throughout the survey process [17].

Representative surveys of the general population are commonly based on sampling frames constructed from lists of addresses or telephone numbers. If the goal of a research team is to estimate SARS-CoV-2 seroprevalence within a population residing in a single municipality, a representative estimate can be obtained using household sampling methods. Frasier et al [18] propose a design for representative COVID-19 seroprevalence studies in the United States using address-based sampling (ABS) methods. With ABS, samples are randomly selected from lists of mailing addresses derived from the US Postal Services' database [19]. Participants are recruited by mail or in-person for study participation, and testing is conducted in neighboring clinics or using self-administered test kits with at-home collection [18]. The ABS methodology has been validated in numerous settings and geographies to have high coverage of the general population [19–21]. Because selection is random and not driven by the participant or the researcher, selection bias due to the sampling method is eliminated. Guidance for sample size determination [18, 22] and the logistics of conducting seroprevalence studies using ABS [18] are available. Sampling techniques such as stratification and clustering can facilitate efficient designs, logistic feasibility, and estimation of subpopulations of interest. Designs can incorporate oversampling of at-risk

and vulnerable populations to allow for robust assessments of seroprevalence in these populations. The application of ABS methods to estimate SARS-CoV-2 seroprevalence is new, so characteristics of nonresponse are still unknown. Established methods to enhance community engagement as well as minimize, measure, and adjust for nonparticipation within the sample [17, 23] should be used to ensure that those who participate are representative of the target population.

Household sampling can be a time- and cost-intensive process. Because of the urgency to obtain seroprevalence estimates quickly, an efficient approach is to partner with an existing study already collecting representative data in the geography of interest. For example, we are working with a local health department that has an established household cohort for estimating factors related to population health in the county. The sample for the cohort has been selected, and participants have already been recruited to complete annual surveys. Representative estimates of seroprevalence will be obtained relatively quickly by recruiting within the existing study sample [24]. Likewise, researchers in Switzerland obtained representative estimates of seroprevalence by sampling former participants from a representative survey of population health [25]. Collaboration between researchers across disciplines and institutions can facilitate these types of timely but representative estimates of seroprevalence.

CONCLUSIONS

Although they are more time-consuming and resource-intensive, representative samples are urgently needed to quantify seroprevalence of COVID-19 and to monitor disease trends over time. These studies will also serve as benchmarks for evaluating the performance of less rigorous methodologies, including convenience samples. Not all researchers can use probability-based sampling methods due to time and cost constraints. In these circumstances, estimates based on extrapolation from convenience samples should clearly outline the assumptions being made, and, when possible, results should be compared with benchmarks from probability-based studies. Although representative studies will take time, we caution against overinterpreting the results of convenience samples in the interim.

Notes

Acknowledgments. We thank the anonymous reviewers for valuable comments that strengthened this manuscript.

Financial support. Funding for seroprevalence studies was provided by the North Carolina Department of Health and Human Services, Carolina Population Center under grant number: NIH P2C HD050924.

Potential conflicts of interest. All authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

References

1. Whitehouse.gov and the Centers for Disease Control and Prevention. Guidelines: opening up America again. Available at: https://www.whitehouse.gov/openingamerica. Accessed 9 June 2020.
2. Akpan N. Here's how to stop the virus from winning. National Geographic. Available at: https://www.nationalgeographic.com/science/2020/06/how-to-stop-coronavirus-surges-from-winning-the-war-cvd/. Accessed 1 July 2020.
3. Winter AK, Hegde ST. The important role of serology for COVID-19 control. Lancet Infect Dis **2020**; 20:758–9.
4. Song J-Y, Yun J-G, Noh J-Y, Cheong H-J, Kim W-J. Covid-19 in South Korea—challenges of subclinical manifestations. N Engl J Med **2020**; 382:1858–9.
5. Mallapaty S. Antibody tests suggest that coronavirus infections vastly exceed official counts. Available at: https://www.nature.com/articles/d41586-020-01095-0. Accessed 9 June 2020.
6. Petherick A. Developing antibody tests for SARS-CoV-2. Lancet **2020**; 395:1101–2.
7. Goodman JD, Rothfeld M. *1 in 5 New Yorkers May Have Had Covid-19, Antibody Tests Suggest*. Available at: https://www.nytimes.com/2020/04/23/nyregion/coronavirus-antibodies-test-ny.html?action=click&module=Spotlight&pgtype=Homepage. Accessed 9 June 2020.
8. Bendavid E, Mulaney B, Sood N, et al. COVID-19 antibody seroprevalence in Santa Clara County, California. medRxiv [Preprint] **2020**: Available at: https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v2
9. Sood N, Simon P, Ebner P, et al. Seroprevalence of SARS-CoV-2–specific antibodies among adults in Los Angeles County, California, on April 10–11, 2020. JAMA **2020**; 323:2425–7.
10. Rosenberg ES, Tesoriero JM, Rosenthal EM, et al. Cumulative incidence and diagnosis of SARS-CoV-2 infection in New York [published online ahead of print June 17, 2020]. Ann Epidemiol **2020**; S1047-2797(20)30201-5. doi:10.1016/j.annepidem.2020.06.004.
11. Yancy CW. COVID-19 and African Americans. JAMA **2020**; 323:1891–2.
12. Larremore DB, Fosdick BK, Bubar KM, et al. Estimating SARS-CoV-2 seroprevalence and epidemiological parameters with uncertainty from serological surveys. medRxiv [Preprint] **2020**: Available at: https://www.medrxiv.org/content/10.1101/2020.04.15.20067066v2.
13. Bosman J, Mervosh S. As virus surges, younger people account for "Disturbing" number of cases. Available at: https://www.nytimes.com/2020/06/25/us/coronavirus-cases-young-people.html. Accessed 2 July 2020.
14. Baker MG, Peckham TK, Seixas NS. Estimating the burden of United

States workers exposed to infection or disease: a key factor in containing risk of COVID-19 infection. PloS One **2020**; 15:e0232452.

15. Sen-Crowe B, McKenney M, Elkbuli A. Social distancing during the COVID-19 pandemic: staying home save lives. Am J Emerg Med **2020**; 38:1519–20.
16. Kolata G. *Coronavirus Infections May Not Be Uncommon, Tests Suggest*. Available at: https://www.nytimes.com/2020/04/21/health/coronavirus-antibodies-california.html. Accessed 12 June 2020.
17. Groves RM, Fowler FJ Jr, Couper MP, Lepkowski JM, Singer E, Tourangeau R. Survey Methodology. Hoboken, NJ: John Wiley & Sons, **2009**.
18. Frasier A, Guyer H, DiGrande L, Domanico R, Cooney D, Eckman S. Design for a mail survey to determine prevalence of SARS-CoV-2 antibodies in the United States. Surv Res Methods **2020**; 14:131–9.
19. Iannacchione VG. The changing role of address-based sampling in survey research. Public Opin Q **2011**; 75:556–75.
20. Shook-Sa BE, Currivan DB, McMichael JP, Iannacchione VG. Extending the coverage of address-based sampling frames: beyond the USPS computerized delivery sequence file. Public Opin Q **2013**; 77:994–1005.
21. Battaglia MP, Dillman DA, Frankel MR, et al. Sampling, data collection, and weighting procedures for address-based sample surveys. J Surv Stat Methodol **2016**; 4:476–500.
22. Schnell R, Smid M. Methodological problems and solutions for sampling in epidemiological SARS-CoV-2 research. Surv Res Methods **2020**; 14:123–9.
23. Holzer JK, Ellis L, Merritt MW. Why we need community engagement in medical research. J Invest Med **2014**; 62:851–5.
24. UNC Gillings School of Global Public Health Communications. Gillings School partners with state, local agencies in North Carolina to study COVID-19 cases with mild or no symptoms. Available at: https://sph.unc.edu/sph-news/gillings-school-partners-with-state-local-agencies-in-north-carolina-to-study-covid-19-cases-with-mild-or-no-symptoms/. Accessed 9 June 2020.
25. Stringhini S, Wisniak A, Piumatti G, et al. Seroprevalence of anti-SARS-CoV-2 IgG antibodies in Geneva, Switzerland (SEROCoV-POP): a population-based study. Lancet **2020**. doi:10.1016/S0140-6736(20)31304-0